Improvement and Assessment of Spectro-Temporal Modulation Analysis for Speech Intelligibility Estimation

Edraki, Amin ; Chan, Wai Yip Geoffrey; Jensen, Jesper; Fogerty, Daniel

Published in:
Interspeech 2019

DOI (link to publication from Publisher):
10.21437/Interspeech.2019-2898

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improvement and Assessment of Spectro-Temporal Modulation Analysis for Speech Intelligibility Estimation

Amin Edraki[1], *Wai-Yip Chan*[1], *Jesper Jensen*[2,3], *Daniel Fogerty*[4]

[1]Department of Electrical and Computer Engineering, Queen's University, Canada
[2]Department of Electronic Systems, Aalborg University, Denmark
[3]Oticon, Denmark
[4]Department of Communication Sciences and Disorders, University of South Carolina, USA

a.erdaki@queensu.ca, chan@queensu.ca, jje@es.aau.dk, fogerty@sc.edu

Abstract

Several recent high-performing intelligibility estimators of acoustically degraded speech signals employ temporal modulation analysis. In this paper, we investigate the utility of using both spectro- and temporal-modulation for estimating speech intelligibility. We modified a pre-existing speech intelligibility estimation scheme (STMI) that was inspired by human auditory spectro-temporal modulation analysis. We produced several variants of the modified STMI and assessed their intelligibility prediction accuracy, in comparison with several high-performing estimators. Among the estimators tested, one of the STMI variants and eSTOI performed consistently well on both noisy and reverberated speech. These results suggest that spectro-temporal modulation analysis is useful for certain degradation conditions such as modulated noise and reverberation.

Index Terms: speech intelligibility, speech quality model, spectro-temporal modulation

1. Introduction

Speech signals captured by human and machine receivers are often distorted by noise and reverberation. Furthermore, while noise reduction algorithms reduce the influence of noise, other degradations may be introduced due to processing. Therefore, the perceptual evaluation of these signals is of a great importance. However, subjective listening tests are costly and time-consuming. Hence, accurate, objective tests can play an important role in the development and evaluation phase for speech processing systems. Among existing successful predictors are several that are based on temporal-modulation analysis. For example, *speech transmission index* (STI) [1] calculates a speech intelligibility (SI) measure based on the reduction in the temporal modulation depth of reverberant or noisy speech. The *short-time objective intelligibility* (STOI) [2] computes a similarity measure between clean and degraded speech temporal modulation envelopes using cross correlation. Both STI and STOI have shown high correlation with SI in reverberation and noisy conditions respectively. *Speech to reverberation modulation energy ratio* (SRMR) [3], a successful SI measure for reverberant and dereverberated speech, relies on analyzing the temporal modulation spectrum of the speech signal. In this study, we employ a biologically inspired spectro-temporal modulation feature-extraction framework for SI estimation. We modify a pre-existing measure within this framework to produce several variants and assess their performance along with other state-of-the-art objective intelligibility measures.

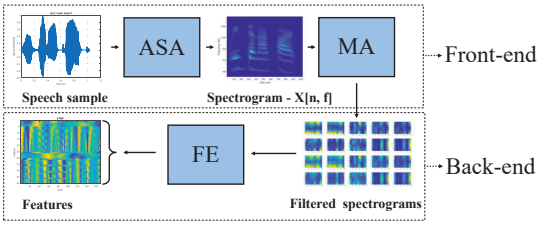

Figure 1: *Spectro-temporal modulation feature (STMF) extraction scheme*

2. Spectro-temporal modulation analysis

Temporal modulation analysis has long been exploited to perform various speech processing tasks such as speech perception modelling [4] and intelligibility prediction [2] [5] [6] [7]. However, emerging auditory psychophysics and neuroscience results suggest that combined spectro-temporal modulation analysis, can better explain human perception [8]. Figure 1 shows a computational model that attempts to mimic salient steps of the auditory system [8]. We use this model to extract spectro-temporal modulation features (STMFs) for estimating SI. STMF extraction comprises three stages of processing:

- Auditory-Spectral Analysis (ASA): This stage mimics the early stages in the human auditory system which transform sound into neural activity patterns represented by acoustic spectra.
- Modulation Analysis (MA): This block mimics the complex set of behaviour associated with neural responses in the primary auditory cortex (A1). These properties can be captured through a bank of modulation analysis (MA) filters associated with important spectro-temporal modulation frequencies (spectro-temporal receptive fields - STRFs).
- Feature Extraction (FE): In this stage, the spectro-temporal decomposition from the MA block is processed to extract a robust set of features for estimating SI.

The following subsections present the computational details of the three stages for our task at hand, namely SI prediction.

2.1. Auditory-spectral and modulation analysis

In the following subsections, we present two combined ASA and MA front-ends used in this paper.

2.1.1. Maryland front-end

A biologically faithful ASA and MA scheme was laid out in [9] [10] [8] (we will denote it as MFN short for Maryland front-

http://dx.doi.org/10.21437/Interspeech.2019-2898

end). Briefly, to construct audio spectrograms, MFN uses 120 overlapping bandpass filters equally spaced over a 5 octave frequency range (24 filters per octave) to model the filtering properties of the cochlea. The output of each filter is processed through a high-pass filter, a non-linear compression and a low-pass filter to mimic hair cell processing. The calculated spectrogram ($X[n, f]$ in Eq. 1) is then processed through a spectro-temporal modulation filter bank (the MA block in Fig. 1) to produce a time-frequency representation of the speech sample. This time-frequency response representation can be expressed as:

$$y[n, f; s, r] = X[n, f] * g[n, f; s, r] \quad (1)$$

where f denotes the acoustic frequency, n is the time, X represents the auditory spectrogram of the signal, and g is the impulse response of a modulation filter tuned to the center spectral modulation frequency s and the temporal modulation frequency r. The filters $g[\cdot]$ are further characterized by their bandwidth and filter dynamics. Multiple spectro-temporal modulation patterns can be constructed by choosing the center spectro-temporal modulation frequencies to form a filter bank that covers important spectro-temporal modulation frequencies in speech perception. For example, Figure 2 shows the impulse response of two two-dimensional bandpass filters, $g_1[\cdot]$ and $g_2[\cdot]$ along with the resultant filtered spectrograms.

In MFN, each Gabor-like modulation filter ($g[\cdot]$ in Eq. 1) is represented as the product of two separate spectral and temporal functions:

$$g[n, f; s, r] = RF[f; s] \cdot h[n; r] \quad (2)$$

where RF is a response field along the acoustic frequency axis and h is a temporal impulse response. A Gabor-like function is used to model the response field function. For a detailed description of the filter parameters, please refer to [8].

2.1.2. *Oldenburg front-end*

In the Oldenburg front-end (OFN) [11], a Mel-spectrogram is calculated from the speech signal using an implementation of the ETSI Distributed Speech Recognition Standard [12]. Eq. 3 through 5 describe the one-dimensional Gabor filters used to construct the two-dimensional modulation filters in this study [11].

$$h_b(x) = \begin{cases} 0.5 - 0.5\cos(\frac{2\pi x}{b}), & -\frac{b}{2} < x < \frac{b}{2}, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$s_\omega(x) = \exp(i\omega x) \quad (4)$$

$$g_{1D}(x) = s_\omega(x)\, h_{\nu/2\omega}(x) \quad (5)$$

where h_b is a Hann envelope function of width b (which is inversely proportional to the number of half-waves under the envelope), s_ω is a sinusoidal function with radian frequency ω and g_{1D} is the product of both. Each two-dimensional spectro-temporal modulation Gabor filter can be constructed as the outer product of two one-dimensional Gabor filters. As suggested in [11], the imaginary and real parts of the Gabor filters enable capturing different phases. For maximum robustness in automatic speech recognition (ASR), the real and imaginary parts were both used to construct 2D separable real filters. Table 1 indicates the set of parameters used in this study. Hence, our spectro-temporal filter bank consists of 25 (5×5) 2D filters [11].

2.2. Feature extraction (FE)

In the following subsections, two feature extraction (FE) schemes are discussed (see Fig. 1). Intelligibility predictions are made based on these features. We denote these operations as "*back-end*" processing.

Table 1: *Parameters used to construct the Gabor filter bank for MA. Scale: spectral modulation frequency. Rate: temporal modulation frequency*

Filter Bank Parameter	**Value(s)**
scale (cycles/channel)	{0, 0.029, 0.060, 0.122, 0.25}
rate (Hz)	{0, 6.2, 9.9, 15.7, 25.0}
ν	3.5
b^{max} (frames)	40

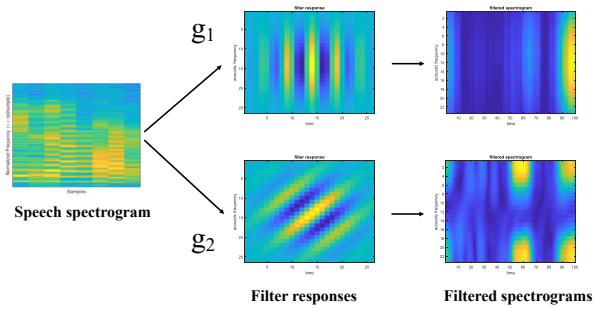

Figure 2: *Sample Gabor-like filter responses along with filtered spectrograms*

2.2.1. *Maryland back-end*

In [13], a FE scheme which we denote as MBN short for Maryland back-end, is discussed based on estimating the total "*energy*" at each specific spectro-temporal modulation frequency. Towards this end, the 4D modulation filtered spectrogram (Eq. 1) is integrated over the acoustic frequency axis (f) to obtain a 3D representation:

$$F_{MBN}[n; s, r] = \sum_f |y[n, f; s, r]| \quad (6)$$

where F_{MBN} denotes a MBN-extracted STMF and y is the modulation filtered spectrogram. Therefore, each STMF demonstrates the variation of total energy over all acoustic frequencies in a specific spectro-temporal modulation frequency as a function of time.

2.2.2. *Oldenburg back-end*

As discussed in [13], a major simplification in the MBN scheme is the integration of the modulation filtered spectrogram over the acoustic frequency axis to produce STMFs (F_{MBN} in Eq. 6). This approach has the benefit of significantly reducing the dimensionality of the output decomposition, but on the other hand ignores the distinct contribution of different acoustic frequencies in the output. In [14], a feature extraction procedure is proposed to create a feature-set based on subsampling the modulation filtered spectrograms to reduce the redundancy of the adjacent acoustic frequency bins, in contrast to integrating over this axis as is done in MBN (Eq. 6). Specifically, a downsampling procedure is proposed [14][11] to select a group of representative channels from each filtered spectrogram by exploiting the fact that the effect of the noise in intermediate frequency ranges (near 1 kHz) is much more significant compared to the edges, and the fact that the signal in nearby acoustic frequency bins in each spectrogram are highly correlated. Therefore, the Oldenburg feature-set is defined as:

$$F_{OBN}[n; s, r, f] = y[n, f; s, r], \quad f \in RC[s, r] \quad (7)$$

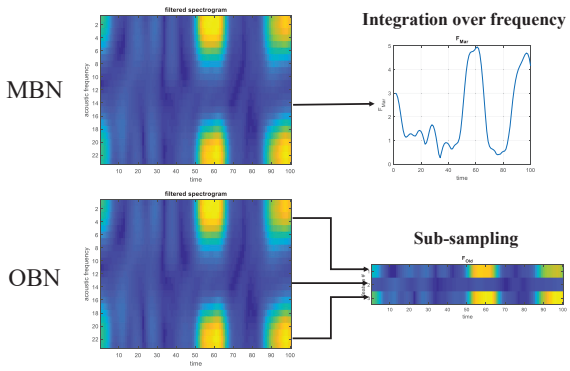

Figure 3: *An example of back-end operations. Top: Maryland back-end (MBN). Bottom: Oldenburg back-end (OBN).*

where $RC[s,r]$ is the set of representative channels for each specific spectral and temporal modulation filtered spectrogram and is a subset of all acoustic frequency channels. Specifically, the center frequency of all filtered spectrograms (corresponding to about 1 kHz) is selected. Moreover, channels with an approximate distance equal to a multiple of 1/4 of the filter width to the center frequency are included. We will denote this scheme by OBN short for Oldenburg back-end. Figure 3 shows the processing steps of two back-ends discussed (MBN and OBN) for a sample speech spectrogram.

2.3. Spectro-temporal modulation index (STMI)

STMI [13] uses the MFN and the MBN described in Sec. 2.1.1 and 2.2.1 to provide a similarity measure between two time-aligned speech samples. One sample serves as a "*clean*" reference, and the other as a degraded signal whose intelligibility is of interest. STMI maps the spectro-temporal modulation contents of the two speech samples to an estimate of SI. STMI computes an intermediate similarity measure between the clean and degraded speech sample as the normalized cross correlation between the clean and degraded STMFs:

$$\rho_0^{STMI}[s,r] = \frac{\langle F^c_{MBN}[n;s,r] - \mu_{F^c_{MBN}}, F^d_{MBN}[n;s,r] - \mu_{F^c_{MBN}} \rangle}{\|F^d_{MBN}[n;s,r] - \mu_{F^c_{MBN}}\| \, \|F^d_{MBN}[n;s,r] - \mu_{F^c_{MBN}}\|} \tag{8}$$

where F^c_{MBN} and F^d_{MBN} denote the STMFs of the clean and degraded speech samples, respectively and μ_F indicates the feature average over time:

$$\mu_F = \mu_F[f,s,r] = \frac{1}{N}\sum_n F[n;f,s,r], \tag{9}$$

where N is the total number of time frames, and the inner product and the induced norm are defined as:

$$\langle H[n], G[n] \rangle = \sum_n H[n]\, G[n], \tag{10}$$

$$\|H[n]\| = \sqrt{\langle H[n], H[n] \rangle}, \tag{11}$$

An overall similarity measure between the clean and degraded time-frequency decompositions is defined as the average over all intermediate similarity measures:

$$\rho^{STMI} = \frac{1}{M}\sum_{s,r} \rho_0^{STMI}[s,r] \tag{12}$$

where M indicates the total number of STMFs (total number of spectro-temporal filters).

2.4. Proposed STMI variants

The original STMI [13] defined by Eq. 12 and is based on MFN and MBN. Here to improve its intelligibility prediction performance, we devised multiple variations of this algorithm using different front-end/back-end combinations. Table 2 summarizes the STMI variations. We can express the intermediate similarity measures for ${}^{O}STMI^{O}$ and ${}^{O}STMI$ by replacing F_{MBN} in Eq. 8 with F_{OBN} (from Eq. 7) and y (from Eq. 1) respectively.

Table 2: *STMI variations based on different front-end/back-end combinations.*

STMI variation	FE/BE combination
${}^{M}STMI^{M}$	MFN and MBN*
${}^{O}STMI^{M}$	OFN and MBN
${}^{O}STMI^{O}$	OFN and OBN
${}^{O}STMI$	OFN and no FE**

**This is the original STMI as proposed in [13]. **In this variant of STMI, the raw 4D filtered spectrgorams ($y[\cdot]$ in Eq. 1) are used as the feature-set.*

Specifically, ${}^{O}STMI^{M}$ uses OFN to generate 4D filtered spectrograms and then integrates over the acoustic frequency axis (MBN) to generate STMFs, and then computes the similarity measure based on this feature-set. ${}^{O}STMI^{O}$ uses the OFN and applies the down-sampling procedure discussed in Sec. 2.2.2 (OBN) to create the feature-set for computing the similarity measure, and ${}^{O}STMI$ uses the OFE and applies no FE to the 4D spectrograms.

3. Data and algorithms

3.1. Datasets

All the noise corrupted speech datasets were collected in previous studies based on measuring the speech recognition of normal hearing adults. We created the reverberation datasets for this study.

3.1.1. Kjems dataset

150 sentences from the Dantale II corpus were degraded by four types of noise: speech-shaped noise (SSN), cafeteria noise, car interior noise and noise from a bottling hall [15]. An ideal binary mask (IBM) and a target binary mask (TBM) were separately applied to the noisy speech, yielding seven categories (IBM and TBM are equivalent in SSN). The noises were mixed with clean speech at three different SNRs to result in 20% speech reception threshold (SRT), 50% SRT, and -60 dB SNR. Eight different relative criteria (RC; local criterion – SNR) were used, including the unprocessed situation, to calculate each binary mask. This resulted in a total of 7 (binary mask/noise categories) x 3 (SNRs) x 8 (RCs) = 168 total conditions.

3.1.2. Noise-reduction dataset I (NR-I)

IEEE sentences and isolated consonants were corrupted by four types of noise (babble, car, street and train) and presented at two SNRs (0 and 5 dB) [16]. These mixtures were then processed by eight speech enhancement methods (spectral subtractive, subspace, statistical model based and Wiener-type algorithms).

3.1.3. Noise-reduction dataset II (NR-II)

Noisy Hagerman sentences (Dutch version) were processed by three non-linear single-microphone noise reduction algorithms

that aimed at finding binary or soft minimum mean-square error estimates of the short-time spectral amplitude. SSN was presented at -8, -6, -4, -2 and 0 dB SNR. This resulted in a total of 20 experimental conditions: 4 processing conditions (unprocessed and three with noise reduction) x 5 SNRs [17].

3.1.4. Modulated noise dataset (Mod noise)

Dantale II sentences were presented in ten different types of modulated noise at six different SNRs. SNRs were selected to cover the full range of performance from completely intelligible to completely unintelligible across the different noise types. Three trials per condition were presented to participants yielding a total of 180 sentences [17].

3.1.5. Reverberation datasets

HINT and IEEE sentences were processed according to 12 reverberation conditions (RCs) with 4 values of reverberation time (0.9, 1.2, 1.5, 2.1 seconds) at each of 3 direct-to-reverberant ratios (0, -10, -20 dB). The room impulse responses were simulated using the image method and convolved with clean utterances to produce reverberant speech at each RC.

3.2. Baseline algorithms

The performance of the STMI variants is compared to several state-of-the-art modulation based SI predictors: *STOI*, an intrusive SI measure based on comparing the temporal envelopes of the clean and degraded signals [2]; *eSTOI*, STOI extended to work for a larger range of degradation conditions [17], and *SRMR*, a non-intrusive SI measure based on analyzing the temporal modulation spectra of degraded speech signals [3] [18]. Both STOI and eSTOI have shown high correlation with SI in noisy conditions. SRMR has shown a high correlation with SI in the presence of reverberation.

4. Results and discussion

Tables 3 and 4 indicate the performance of different intelligibility predictors in terms of the Pearson correlation coefficient ρ_p and Spearman correlation coefficient ρ_s, respectively. STOI compares the temporal modulation envelopes of the clean and degraded signals by means of their cross correlation. STOI shows high correlation with the intelligibility of degraded speech in all degradation conditions except modulated noise which is in accordance with the results reported in [17] as well as [19]. All the variations of STMI outperform STOI in this condition; in case the of ${}^{O}STMI^{O}$ by a huge margin. This suggests the significance of spectral modulation analysis in the presence of modulated noise, as STOI only performs temporal modulation analysis.

${}^{M}STMI^{M}$, the original STMI, performs poorly with noisy speech. All three extended STMI variants show much improved performance. We attribute the improvement to replacing MFN with OFN. Moreover, the fact that ${}^{O}STMI$ outperforms ${}^{O}STMI^{M}$ in almost all degradation conditions suggests that integration over the acoustic frequency axis as done in MBN detrimentally removes the distinct effects of different frequency channels on SI prediction.The feature extraction/selection method of OBN improves the overall performance of STMI significantly as demonstrated by ${}^{O}STMI^{O}$. The subsampling procedure introduced in OBN selects most of the STMFs from the filtered spectrograms with high spectral modulation frequencies. This might imply the importance of higher spectral modulation frequencies in SI prediction. The strong performance of ${}^{O}STMI^{O}$ is overmatched by eSTOI. The subspace decomposition of temporal modulation envelopes across acoustic frequencies performed in eSTOI effects the exploitation of spectral modulation information. Hence, both ${}^{O}STMI^{O}$ and eSTOI suggest the usefulness of spectro-temporal modulation information in SI prediction. Recent work [14] [11] makes a similar observation but for ASR systems. Arguably, STMI provides a more faithful representation of the processing in the human auditory system compared to eSTOI.

As the results in Table 3 and 4 indicate, STOI, eSTOI and STMI all show high correlation with SI in the reverberation condition; this contradicts the finding in [20] for STOI. Despite using no reference (non-intrusive), SRMR performs well for reverberant though not for noisy speech.

Table 3: *Performance of SI predictors in terms of Pearson correlation coefficient ρ_p between subjective intelligiblity and algorithm output*

	Kjems	NR I	NR II	Mod noise	HINT rev	IEEE rev
STOI	0.87	0.86	0.97	0.45	0.91	0.91
eSTOI	**0.89**	**0.90**	**0.97**	0.85	0.92	0.90
${}^{M}STMI^{M}$	0.70	0.64	0.56	0.50	0.85	0.92
${}^{O}STMI^{M}$	0.74	0.81	0.93	0.75	0.88	0.90
${}^{O}STMI$	0.80	0.86	0.89	0.81	0.89	0.93
${}^{O}STMI^{O}$	0.87	0.89	0.89	**0.88**	**0.96**	**0.95**
SRMR	0.12	0.37	0.42	0.24	0.84	0.90

Table 4: *Performance of SI predictors in terms of Spearman correlation coefficient ρ_s between subjective intelligiblity and algorithm output*

	Kjems	NR I	NR II	Mod noise	HINT rev	IEEE rev
STOI	0.91	0.80	0.96	0.48	**0.97**	**0.96**
eSTOI	**0.91**	**0.87**	**0.98**	**0.92**	0.96	**0.96**
${}^{M}STMI^{M}$	0.70	0.65	0.58	0.51	0.90	0.94
${}^{O}STMI^{M}$	0.77	0.74	0.96	0.81	0.90	0.94
${}^{O}STMI$	0.85	0.80	0.88	0.87	0.91	**0.96**
${}^{O}STMI^{O}$	0.90	0.84	0.90	0.91	**0.97**	**0.96**
SRMR	0.12	0.37	0.46	0.24	0.86	0.95

5. Conclusion

We sought to determine whether spectral modulation information could be beneficial, in addition to the often-used temporal modulation information, in estimating the intelligibility of acoustically degraded speech signals. Towards this end, we modified STMI, a pre-existing human central auditory processing inspired spectro-temporal modulation analysis scheme. One modified variant, ${}^{O}STMI^{O}$ provided good estimation performance on all datasets tested, matching eSTOI, the top performing baseline scheme. The results suggest that spectral modulation information is useful for SI prediction in modulated noise conditions, an important condition given the prevalence of mobile acoustic interfaces.

6. References

[1] P. CODE, "Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index," 2003.

[2] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted

noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[3] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[4] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[5] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7-8, pp. 678–692, 2010.

[6] S. Jørgensen and T. Dau, "Modeling speech intelligibility based on the signal-to-noise envelope power ratio," 2014.

[7] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, 2013.

[8] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[9] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," Tech. Rep., 1991.

[10] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE transactions on speech and audio processing*, vol. 2, no. 3, pp. 421–435, 1994.

[11] M. R. Schädler and B. Kollmeier, "Separable spectro-temporal gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2047–2059, 2015.

[12] I. O. for Standardization, "Etsi standard 201 108 v1.1.3," *https://www.etsi.org/website/Technologies/Dis- tributed-SpeechRecognition.aspx*, 2003.

[13] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[14] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[15] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1415–1426, 2009.

[16] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[17] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[18] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2014, pp. 55–59.

[19] S. Jørgensen, R. Decorsière, and T. Dau, "Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1401–1410, 2015.

[20] H. R. Iborra, "Analysing the Effects of Auditory Processing and the Decision Metric on Speech Intelligibility Prediction," Master's thesis, DTU - Technical University of Denmark, Kgs. Lyngby, Ørsteds Plads, Building 352, 2800 Kgs. Lyngby, Denmark, 2015.